

# Smoothing with Couplings of Conditional Particle Filters <sup>\*</sup>

Pierre E. Jacob<sup>†</sup>

Department of Statistics, Harvard University

Fredrik Lindsten and Thomas B. Schön

Department of Information Technology, Uppsala University

January 10, 2017

## Abstract

In state space models, smoothing refers to the task of estimating a latent stochastic process given noisy measurements related to the process. We propose the first unbiased estimator of smoothing expectations. The lack-of-bias property has methodological benefits, as it allows for a complete parallelization of the algorithm and for computing accurate confidence intervals. The method combines two recent breakthroughs: the first is a generic debiasing technique for Markov chains due to Rhee and Glynn, and the second is the introduction of a uniformly ergodic Markov chain for smoothing, the conditional particle filter of Andrieu, Doucet and Holenstein. We show how a combination of the two methods delivers practical estimators, upon the introduction of couplings between conditional particle filters. The algorithm is widely applicable, has minimal tuning parameters and is amenable to modern computing hardware. We establish the validity of the proposed estimator under mild assumptions. Numerical experiments illustrate its performance in a toy model and in a Lotka-Volterra model with an intractable transition density.

*Keywords:* common random numbers, couplings, particle filtering, particle smoothing, resampling algorithms.

---

<sup>\*</sup>This research is financially supported by the Swedish Foundation for Strategic Research (SSF) via the project *ASSEMBLE* and the Swedish research Council (VR) via the projects *Learning of complex dynamical systems* (Contract number: 637-2014-466) and *Probabilistic modeling of dynamical systems* (Contract number: 621-2013-5524).

<sup>†</sup>Corresponding author: [pjacob@fas.harvard.edu](mailto:pjacob@fas.harvard.edu). Code available at: [github.com/pierrejacob/](https://github.com/pierrejacob/).

# 1 Introduction

## 1.1 Goal and content

In state space models, the observations are treated as noisy measurements related to an underlying stochastic process. The problem of smoothing refers to the estimation of trajectories of the underlying process given the observations. For finite state spaces and linear Gaussian models, smoothing can be performed analytically. In general models, approximations like the ones offered by Monte Carlo methods are required, and many state-of-the-art methods are based on particle filters (Lindsten and Schön, 2013; Douc et al., 2014; Kantas et al., 2015). We propose a new method for smoothing, applicable for general state space models. Unlike existing methods, the proposed estimators are unbiased, which has direct benefits for parallelization and allows the construction of accurate confidence intervals.

The proposed method combines recently proposed conditional particle filters (Andrieu et al., 2010) with debiasing techniques for Markov chains (Glynn and Rhee, 2014). After introducing the context and notation, we describe an unbiased smoothing estimator in Section 2, which relies on coupled resampling. Coupled resampling schemes are discussed in Section 3. The validity of the proposed smoother is established in Section 4. Performance and variance reduction are discussed in Section 5, and the proposed method is illustrated numerically in Section 6. The appendices contain proofs (Appendix A), additional numerical experiments (Appendix B) and pseudo-code (Appendix C).

## 1.2 Smoothing in state space models

In state space models, a latent stochastic process  $(x_t)_{t \geq 0}$  takes values in  $\mathbb{X} \subset \mathbb{R}^{d_x}$ , and the observations  $(y_t)_{t \geq 1}$  are in  $\mathbb{Y} \subset \mathbb{R}^{d_y}$  for some  $d_x, d_y \in \mathbb{N}$ . A model specifies an initial distribution  $m_0(dx_0|\theta)$  and a transition kernel  $f(dx_t|x_{t-1}, \theta)$  for the latent process. Conditionally upon the latent process, the observations are independent and their distribution is given by a measurement kernel  $g(dy_t|x_t, \theta)$ . The model is parameterized by  $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ , for  $d_\theta \in \mathbb{N}$ . Filtering consists in approximating the distribution  $p(dx_t|y_{1:t}, \theta)$  for all times  $t \geq 1$ , whereas smoothing consists in approximating the distribution  $p(dx_{0:T}|y_{1:T}, \theta)$  for a fixed time horizon  $T$ , where for  $s, t \in \mathbb{N}$ , we write  $s:t$  for the set  $\{s, \dots, t\}$ , and  $v_{s:t}$  for the vector  $(v_s, \dots, v_t)$ . In the following, the parameter  $\theta$  is fixed and removed from the notation. Denote by  $h$  a test function on  $\mathbb{X}^{T+1}$ , of which we want

to compute the expectation with respect to the smoothing distribution  $\pi(dx_{0:T}) = p(dx_{0:T}|y_{1:T})$ ; we write  $\pi(h)$  for  $\int_{\mathbb{X}^{T+1}} h(x_{0:T})\pi(dx_{0:T})$ .

### 1.3 Particle filters

Most Monte Carlo approaches for smoothing rely on particle filters (Gordon et al., 1993; Doucet et al., 2001; Cappé et al., 2005; Doucet and Johansen, 2011). The particle filter generates weighted samples denoted by  $(w_t^k, x_t^k)_{k=1}^N$ , for all  $t \in \mathbb{N}$ , where the particle locations  $(x_t^k)_{k=1}^N$  are samples in  $\mathbb{X}$  and the weights  $(w_t^k)_{k=1}^N$  are non-negative reals summing to one. The number  $N \in \mathbb{N}$  of particles is specified by the user—the computational cost of the algorithm is linear in  $N$ , while the approximation of the filtering distribution  $p(dx_t|y_{1:t})$  by  $\sum_{k=1}^N w_t^k \delta_{x_t^k}(dx_t)$  becomes more precise as  $N$  increases (e.g. Del Moral, 2004).

In the basic, so-called, bootstrap particle filter, random variables are used to initialize, to re-sample and to propagate the particles. Initially, we sample  $x_0^k \sim m_0(dx_0)$  for all  $k \in 1 : N$ , or equivalently, we compute  $x_0^k = M(U_0^k)$  where  $M$  is a (deterministic) function and  $U_0^{1:N}$  random variables. The initial weights  $w_0^k$  are set to  $N^{-1}$ . Consider each step  $t \geq 0$  of the algorithm. In the resampling step, a vector of ancestor variables  $a_t^{1:N} \in \{1, \dots, N\}^N$  is sampled from some distribution  $r$ :  $a_t^{1:N} \sim r(da^{1:N}|w_t^{1:N})$ . The propagation step consists in drawing  $x_{t+1}^k \sim f(dx_{t+1}|x_t^{a_t^k})$ , or equivalently, computing  $x_{t+1}^k = F(x_t^{a_t^k}, U_{t+1}^k)$ , where  $F$  is a function and  $U_{t+1}^{1:N}$  random variables. The next weights are computed as  $w_{t+1}^k \propto g(y_{t+1}|x_{t+1}^k)$ , then normalized to sum to one; and the algorithm proceeds. The resampling distribution  $r$  is an algorithmic choice; a standard condition for its validity is that, under  $r$ ,  $\mathbb{P}(a_t^k = j) = w_t^j$  for all  $j, k$ ; various schemes satisfy this condition (e.g. Douc and Cappé, 2005; Murray et al., 2016). Pseudo-code for the particle filter is provided in Appendix C. We will use bold fonts to denote vectors of objects indexed by  $k \in 1 : N$ , for instance  $(\mathbf{w}_t, \mathbf{x}_t) = (w_t^k, x_t^k)_{k=1}^N$  or  $\mathbf{U}_t = U_t^{1:N}$ .

### 1.4 Particle smoothers

One can store the generated paths (Jacob et al., 2015), denoted by  $(x_{0:T}^k)_{k=1}^N$ , at the final time  $T$ . The path  $x_{0:T}^k$  is obtained by tracing back the lineage of the latest particle  $x_T^k$ : define  $b_T^k = k$  and  $b_t^k = a_t^{b_{t+1}^k}$  for all  $0 \leq t < T$ . Then the  $t$ -th component of  $x_{0:T}^k$  is  $x_t^{b_t^k}$ . The approximation of the smoothing distribution  $\pi(dx_{0:T})$  by  $\sum_{k=1}^N w_T^k \delta_{x_{0:T}^k}(dx_{0:T})$  becomes more precise as  $N$  goes to infinity, for fixed  $T$ ; however it is known to degenerate quickly with respect to  $T$ , so that the

bootstrap particle filter is rarely used for smoothing. Popular smoothing methods include the fixed-lag smoother, which approximates marginally  $p(dx_t|y_{0:T})$  by  $p(dx_t|y_{0:t+L})$  for a small value  $L$ , and then approximates  $p(dx_t|y_{0:t+L})$  with a bootstrap particle filter. More sophisticated particle smoothers include the forward filtering backward smoother (FFBS), which involves a backward pass that reweights the particles obtained during the filtering pass (Doucet et al., 2000).

## 1.5 Conditional particle filters

The conditional particle filter (CPF) is a Markov kernel leaving the smoothing distribution invariant (Andrieu et al., 2010); extensions include backward sampling (Whiteley, 2010) and ancestor sampling (Lindsten et al., 2014) (see also Chopin and Singh, 2015; Andrieu et al., 2013; Lindsten et al., 2015). Given a reference trajectory  $X = x_{0:T}$ , the CPF generates a new trajectory  $X' = x'_{0:T}$  as follows.

At the initial step, we draw  $\mathbf{U}_0$  and compute  $x_0^k = M(U_0^k)$  for all  $k \in 1 : N - 1$ ; we set  $x_0^N = x_0$ , and  $w_0^k = N^{-1}$  for all  $k$ . At each step  $t$ , we draw  $a_t^{1:N-1} \sim r(da^{1:N-1}|w_t^{1:N})$  from e.g. a categorical distribution, and set  $a_t^N = N$ ; other resampling schemes can be implemented (Chopin and Singh, 2015). In the propagation step, we draw  $\mathbf{U}_{t+1}$  and compute  $x_{t+1}^k = F(x_t^{a_t^k}, U_{t+1}^k)$  for  $k \in 1 : N - 1$ ; we set  $x_{t+1}^N = x_{t+1}$ . The weighting step computes  $w_{t+1}^k \propto g(y_{t+1}|x_{t+1}^k)$ , for all  $k \in 1 : N$ . The algorithm guarantees that the reference trajectory  $x_{0:T}$  is among the  $N$  final trajectories. At the final step, we draw  $b_T$  with probabilities  $\mathbf{w}_T$  and retrieve the corresponding trajectory  $x_{0:T}^{b_T}$ , which is the output trajectory  $X'$ . We write  $X' \sim \text{CPF}(X, \cdot)$ . Iterating the CPF kernel, we construct a chain  $(X^{(n)})_{n \geq 0}$  on the space  $\mathbb{X}^{T+1}$  of trajectories, which admits the smoothing distribution  $\pi(dx_{0:T})$  as invariant distribution (Andrieu et al., 2010). Pseudo-code for the CPF is provided in Appendix C.

One can directly use the chain  $(X^{(n)})_{n \geq 0}$  to approximate smoothing expectations  $\pi(h)$ , using ergodic averages. However, this raises the usual challenging questions of parallelization and convergence diagnostics for MCMC. Our proposed method relies on the CPF kernel, but is *not* an MCMC method. It is fully parallelizable and does not require burn-in or convergence diagnostics, thanks to the lack-of-bias property.

## 2 Unbiased smoothing

### 2.1 Debiassing conditional particle filters

A technique to turn Markov kernels into unbiased estimators has been proposed by Glynn and Rhee (2014). That work follows a series of unbiased estimation techniques in varied settings (e.g. Rhee and Glynn, 2012; Vihola, 2015, and references therein). We describe the idea informally, in the specific setting of smoothing with conditional particle filters; mathematical rigor is deferred to Section 4. Denote by  $(X^{(n)})_{n \geq 0}$  and  $(\tilde{X}^{(n)})_{n \geq 0}$  two Markov chains with invariant distribution  $\pi$ . Assume that, for all  $n \geq 0$ ,  $X^{(n)}$  and  $\tilde{X}^{(n)}$  have the same marginal distribution, and that  $\lim_{n \rightarrow \infty} \mathbb{E}[h(X^{(n)})] = \pi(h)$ . Writing the limit as a telescopic sum, and then assuming that we can swap the infinite sum and the expectation, we have

$$\pi(h) = \mathbb{E}[h(X^{(0)})] + \sum_{n=1}^{\infty} \mathbb{E}[h(X^{(n)}) - h(\tilde{X}^{(n-1)})] = \mathbb{E}[h(X^{(0)}) + \sum_{n=1}^{\infty} h(X^{(n)}) - h(\tilde{X}^{(n-1)})].$$

Then the random variable  $H$ , defined as  $H = h(X^{(0)}) + \sum_{n=1}^{\infty} h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ , is an unbiased estimator of  $\pi(h)$ . Assume further that there exists a time  $\tau$ , termed the *meeting time*, such that  $X^{(n)} = \tilde{X}^{(n-1)}$  almost surely for all  $n \geq \tau$ . Then  $H$  only involves finitely many non-zero terms:

$$H = h(X^{(0)}) + \sum_{n=1}^{\tau-1} h(X^{(n)}) - h(\tilde{X}^{(n-1)}). \quad (1)$$

We refer to  $H$  as the Rhee–Glynn smoothing estimator. We will construct the chains in such a way that the meeting time  $\tau = \inf\{n \geq 0 : X^{(n)} = \tilde{X}^{(n-1)}\}$  has a finite expectation.

### 2.2 Coupled conditional particle filters

To generate chains  $(X^{(n)})_{n \geq 0}$  and  $(\tilde{X}^{(n)})_{n \geq 0}$  amenable to the Rhee–Glynn estimator, we introduce coupled conditional particle filters (CCPF). The CCPF is a Markov kernel on the space of pairs of trajectories, which follows the CPF procedure for each path, using common random numbers. Two particle systems are created, denoted by  $(\mathbf{w}_t, \mathbf{x}_t)$  and  $(\tilde{\mathbf{w}}_t, \tilde{\mathbf{x}}_t)$  at each step  $t$ . We denote the reference trajectories by  $X = x_{0:T}$  and  $\tilde{X} = \tilde{x}_{0:T}$ , and describe how two new trajectories are produced.

Initially, we draw  $\mathbf{U}_0 = U_0^{1:N}$  and compute  $x_0^k = \tilde{x}_0^k = M(U_0^k)$  for all  $k \in 1 : N - 1$ ; we set  $x_0^N = x_0$ ,  $\tilde{x}_0^N = \tilde{x}_0$ , and  $\tilde{w}_0^k = w_0^k = N^{-1}$  for all  $k$ . At this stage,  $N - 1$  particles are identical

across both particle systems. At each step  $t$ , we draw ancestor indices such that, marginally,  $a_t^{1:N-1} \sim r(da^{1:N-1}|w_t^{1:N})$  and  $\tilde{a}_t^{1:N-1} \sim r(d\tilde{a}^{1:N-1}|\tilde{w}_t^{1:N})$ ; and we set  $\tilde{a}_t^N = a_t^N = N$ . How this resampling step should be performed will be elaborated in Section 3; we can simply think of this step as performing multinomial resampling twice, with common random numbers. In the propagation step, we draw  $\mathbf{U}_{t+1}$  and compute  $x_{t+1}^k = F(x_t^{a_t^k}, U_{t+1}^k)$  and  $\tilde{x}_{t+1}^k = F(\tilde{x}_t^{\tilde{a}_t^k}, U_{t+1}^k)$  for  $k \in 1 : N - 1$ . We set  $x_{t+1}^N = x_{t+1}$  and  $\tilde{x}_{t+1}^N = \tilde{x}_{t+1}$ . The weighting step computes  $w_{t+1}^k \propto g(y_{t+1}|x_{t+1}^k)$  and  $\tilde{w}_{t+1}^k \propto g(y_{t+1}|\tilde{x}_{t+1}^k)$ , for all  $k \in 1 : N$ .

At the final step, we draw  $\tilde{b}_T$  and  $b_T$  with probabilities  $\mathbf{w}_T$  and  $\tilde{\mathbf{w}}_T$  respectively, using a common random number. We retrieve the corresponding trajectories  $x_{0:T}^{b_T}$  and  $\tilde{x}_{0:T}^{\tilde{b}_T}$ , which we denote by  $X'$  and  $\tilde{X}'$ . This completes the description of the coupled conditional particle filter, and we write  $(X', \tilde{X}') \sim \text{CCPF}((X, \tilde{X}), \cdot)$ . Pseudo-code for the CC PF is provided in Appendix C. By construction, it is marginally equivalent to a CPF kernel. Jointly, the CC PF kernel allows for the possibility of outputting a pair of identical trajectories. Indeed, from  $N - 1$  identical particles at step 0, a number might keep on being identical at step 1, i.e. those for which  $a_0^k = \tilde{a}_0^k$ ; and so forth up to the final step  $T$ . Thus, although it is potentially very small, especially for large  $T$ , the probability of the event  $\{X' = \tilde{X}'\}$  is non-zero.

### 2.3 Rhee–Glynn smoothing estimator

We start from two trajectories  $X^{(0)}$  and  $\tilde{X}^{(0)}$ , generated as follows. We run two independent particle filters, producing two sets of paths,  $(\mathbf{w}_T, \mathbf{x}_{0:T})$  and  $(\tilde{\mathbf{w}}_T, \tilde{\mathbf{x}}_{0:T})$ . We draw one path from each set, according to the probabilities  $\mathbf{w}_T$  and  $\tilde{\mathbf{w}}_T$  respectively; this is written  $X^{(0)} \sim \text{PF}$  and  $\tilde{X}^{(0)} \sim \text{PF}$ . We then perform one step of the CPF given  $X^{(0)}$ , yielding  $X^{(1)}$ . Going forward, we use the CC PF kernel to obtain the pair of trajectories  $(X^{(n)}, \tilde{X}^{(n-1)})$  given  $(X^{(n-1)}, \tilde{X}^{(n-2)})$ , for all  $n \geq 2$ . We summarize the procedure to compute the unbiased estimator  $H$  in Algorithm 1.

To estimate the smoothing functional  $\pi(h)$ , we propose to sample  $R$  Rhee–Glynn estimators independently, denoted by  $H^{(r)}$  for  $r \in 1 : R$ , and to take the empirical average  $\bar{H} = R^{-1} \sum_{r=1}^R H^{(r)}$ . That average is unbiased and converges to  $\pi(h)$  at the standard Monte Carlo rate as  $R \rightarrow \infty$ . The form of  $\bar{H}$  as an average of independent unbiased estimators has two practical benefits.

1. Complete parallelization of the computation of the terms  $H^{(r)}$  is possible. On the contrary, particle-based methods are not entirely parallelizable due to the resampling step (Murray et al., 2016; Lee and Whiteley, 2015).

---

**Algorithm 1** Rhee–Glynn smoothing estimator.

---

- Draw  $X^{(0)} \sim \text{PF}$ ,  $\tilde{X}^{(0)} \sim \text{PF}$  (independently) and draw  $X^{(1)} \sim \text{CPF}(X^{(0)}, \cdot)$ .
  - Compute  $\Delta^{(0)} = h(X^{(0)})$  and  $\Delta^{(1)} = h(X^{(1)}) - h(\tilde{X}^{(0)})$ , set  $H = \Delta^{(0)} + \Delta^{(1)}$ .
  - For  $n = 2, 3, \dots$ ,
    - Draw  $(X^{(n)}, \tilde{X}^{(n-1)}) \sim \text{CCPF}((X^{(n-1)}, \tilde{X}^{(n-2)}), \cdot)$ .
    - If  $X^{(n)} = \tilde{X}^{(n-1)}$ , then  $n$  is the meeting time  $\tau$ : exit the loop.
    - Compute  $\Delta^{(n)} = h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ , set  $H \leftarrow H + \Delta^{(n)}$ .
  - Return  $H$ .
- 

2. Using the central limit theorem, we can construct accurate error estimators. Error estimators for particle smoothers have not yet been proposed in the literature.

The only tuning parameter is the number of particles  $N$  used within the CCPF, which implicitly sets the number of independent estimators  $R$  that can be obtained within a fixed computing budget. The cost of producing an unbiased estimator  $H$  is of order  $NT \times \mathbb{E}[\tau]$ , and the expectation of  $\tau$  is seen empirically to decrease with  $N$ , so that the choice of  $N$  is not obvious; in practice we recommend choosing a value of  $N$  large enough so that the meeting time occurs within a few steps, but other considerations such as memory cost could be taken into account. The memory cost for each estimator is of order  $T + N \log N$  in average (Jacob et al., 2015). This memory cost holds also when using ancestor sampling (Lindsten et al., 2014), whereas backward sampling (Whiteley, 2010) results in a memory cost of  $NT$ . As in Glynn and Rhee (2014), we could obtain a central limit theorem parameterized by the computational budget instead of the number of samples (Glynn and Whitt, 1992).

### 3 Coupled resampling

We discuss the coupled resampling step, for which multiple implementations are possible; the temporal index  $t$  is removed from the notation, and we consider the problem of sampling  $a^{1:N-1}$  and  $\tilde{a}^{1:N-1}$  such that  $\mathbb{P}(a = j) = w^j$  and  $\mathbb{P}(\tilde{a} = j) = \tilde{w}^j$  for all  $j \in 1 : N$ . The latter ensures the overall validity of the conditional particle filter (Andrieu et al., 2010).

### 3.1 Sampling pairs of ancestor indices

Instead of seeing the coupled resampling step as a way of sampling two vectors of indices  $a^{1:N-1}$  and  $\tilde{a}^{1:N-1}$ , possibly using common random numbers, we consider it to be the sampling of  $N - 1$  pairs of indices in  $\{1, \dots, N\}$ . A joint distribution on  $\{1, \dots, N\}^2$  is characterized by a matrix  $P$  with non-negative entries  $P^{kj}$ , for  $k, j \in \{1, \dots, N\}$ , that sum to one. The value  $P^{kj}$  represents the probability of the event  $(a, \tilde{a}) = (k, j)$ . We consider the set  $\mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$  of matrices  $P$  such that  $P\mathbf{1} = \mathbf{w}$  and  $P^\top \mathbf{1} = \tilde{\mathbf{w}}$ , where  $\mathbf{1}$  denotes a column vector of  $N$  ones. This ensures  $\mathbb{P}(a = j) = w^j$  and  $\mathbb{P}(\tilde{a} = j) = \tilde{w}^j$  for  $j \in 1 : N$ .

The choice  $P = \mathbf{w} \tilde{\mathbf{w}}^\top$  corresponds to an independent coupling of  $\mathbf{w}$  and  $\tilde{\mathbf{w}}$ . Sampling  $(a, \tilde{a})$  from this matrix  $P$  could be done by sampling  $a$  with probabilities  $\mathbf{w}$  and  $\tilde{a}$  with probabilities  $\tilde{\mathbf{w}}$ , independently. Performing multinomial resampling with common random numbers, as suggested in Section 2.2, is another way of sampling from  $P = \mathbf{w} \tilde{\mathbf{w}}^\top$ , in order  $N$  operations.

Any choice of probability matrix  $P \in \mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$ , and of a way of sampling  $P$ , leads to a coupled resampling scheme that can be used in the CCPF. In turn, this corresponds to a smoothing estimator following Algorithm 1. We describe a choice for  $P \in \mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$ , with the aim of coupling pairs of particle systems while keeping the overall cost of the algorithm linear in  $N$ . Other choices of coupled resampling schemes are given in recent technical reports (Jacob et al., 2016; Sen et al., 2016), following seminal works such as Lee (2008); Pitt (2002). We mention links between coupled resampling and optimal transport: if  $P$  is defined as the optimal transport coupling between weighted empirical measures defined by  $(\mathbf{w}, \mathbf{x})$  and  $(\tilde{\mathbf{w}}, \tilde{\mathbf{x}})$ , then the expected distance between resampled particles  $x^a$  and  $\tilde{x}^{\tilde{a}}$  is minimized over all choices of  $P$ . This approach, however, would be super-linear in  $N$ .

### 3.2 Index-coupled resampling

We consider the *index-coupled* resampling scheme, used by Chopin and Singh (2015) in their theoretical analysis of CPF, and by Jasra et al. (2015) for multilevel Monte Carlo. The idea of index-coupling is to maximize the probability of sampling pairs  $(a, \tilde{a})$  such that  $a = \tilde{a}$ , by computing the matrix  $P \in \mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$  with maximum entries on its diagonal. The scheme is intuitive at the initial step of the CCPF, when  $x_0^k = \tilde{x}_0^k$  for all  $k \in 1 : N - 1$ . At later steps, the number of identical pairs across both filters might be small or null. In any case, at step  $t + 1$ , the same random number  $U_{t+1}^k$  is used to compute  $x_{t+1}^k$  and  $\tilde{x}_{t+1}^k$  from their ancestors. Therefore, by sampling  $a_t^k = \tilde{a}_t^k$ , we



select pairs that were computed with common random numbers at the previous step, and give them common random numbers  $U_{t+1}^k$  again. The scheme maximizes the number of consecutive steps at which common random numbers are given to each pair. We discuss below the appeal of this feature, but we first describe how to implement the scheme linearly in  $N$  (similarly to the construction of maximal couplings, [Lindvall \(2002\)](#)).

For all  $k \in 1 : N$ , the matrix  $P$  has to satisfy  $P^{kk} \leq \min(w^k, \tilde{w}^k)$ , otherwise one of the marginal constraints would be violated. We tentatively write  $P = \alpha \text{diag}(\boldsymbol{\mu}) + (1 - \alpha)R$ , where  $\boldsymbol{\mu} = \boldsymbol{\nu}/\alpha$  with  $\boldsymbol{\nu} = \min(\mathbf{w}, \tilde{\mathbf{w}})$  (element-wise),  $\alpha = \sum_{k=1}^N \nu^k$ , and  $R$  is a residual matrix with zeros on the diagonal. Such a  $P$  has maximum trace among all matrices in  $\mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$ . We now look for  $R$  such that  $P \in \mathcal{J}(\mathbf{w}, \tilde{\mathbf{w}})$  and such that sampling from  $P$  can be done in a linear cost in  $N$ . From the marginal constraints, the matrix  $R$  needs to satisfy, for all  $k \in 1 : N$ ,  $\nu^k + (1 - \alpha) \sum_{j=1}^N R^{kj} = w^k$  and  $\nu^k + (1 - \alpha) \sum_{j=1}^N R^{jk} = \tilde{w}^k$ . Among all the matrices  $R$  that satisfy these constraints, the choice  $R = \mathbf{r}\tilde{\mathbf{r}}^\top$ , where  $\mathbf{r} = (\mathbf{w} - \boldsymbol{\nu})/(1 - \alpha)$  and  $\tilde{\mathbf{r}} = (\tilde{\mathbf{w}} - \boldsymbol{\nu})/(1 - \alpha)$ , is such that we can sample pairs of indices from  $R$  by sampling from  $\mathbf{r}$  and  $\tilde{\mathbf{r}}$  independently, for a linear cost in  $N$ . Thus we define the index-coupled matrix  $P$  as

$$P = \alpha \text{diag}(\boldsymbol{\mu}) + (1 - \alpha) \mathbf{r}\tilde{\mathbf{r}}^\top. \quad (2)$$

Using this mixture representation, sampling from  $P$  can be done linearly in  $N$ , without in fact constructing or storing  $P$  explicitly in memory. First, flip an  $\alpha$ -coin. If it comes out heads, then sample  $a$  according to  $\boldsymbol{\mu}$  and set  $\tilde{a} = a$ , otherwise sample  $a$  according to  $\mathbf{r}$  and  $\tilde{a}$  according to  $\tilde{\mathbf{r}}$ , independently or using common random numbers.

Under assumptions on the model, using common random numbers to propagate a pair of particles will result in that pair getting closer. We can formulate assumptions on the function  $x \mapsto \mathbb{E}[F(x, U)]$ , where the expectation is with respect to  $U$ , assuming e.g. that it is Lipschitz. In an auto-regressive model where  $F(x, U) = \theta x + U$ , where  $\theta \in (-1, 1)$  and  $U$  is the innovation term, we have  $|F(x, U) - F(\tilde{x}, U)| = |\theta||x - \tilde{x}|$ , thus the pair of particles converges geometrically fast. Discussions on the generality of this kind of assumptions for Markov chains can be found in [Diaconis and Freedman \(1999\)](#).

## 4 Theoretical properties

We give three sufficient conditions for the validity of Rhee–Glynn smoothing estimators.

**Assumption 1.** The measurement density of the model is bounded from above:

$$\exists \bar{g} < \infty, \quad \forall y \in \mathbb{Y}, \quad \forall x \in \mathbb{X}, \quad g(y|x) \leq \bar{g}.$$

**Assumption 2.** The resampling probability matrix  $P$ , constructed from the weight vectors  $\mathbf{w}$  and  $\tilde{\mathbf{w}}$ , is such that

$$\forall k \in \{1, \dots, N\}, \quad P^{kk} \geq w^k \tilde{w}^k.$$

Furthermore, if  $\mathbf{w} = \tilde{\mathbf{w}}$ , then  $P$  is a diagonal matrix with entries given by  $\mathbf{w}$ .

**Assumption 3.** Let  $(X^{(n)})_{n \geq 0}$  be a Markov chain generated by the conditional particle filter. The test function  $h$  is such that

$$\mathbb{E} \left[ h(X^{(n)}) \right] \xrightarrow{n \rightarrow \infty} \pi(h).$$

Furthermore, there exists  $\delta > 0$ ,  $n_0 < \infty$  and  $C < \infty$  such that

$$\forall n \geq n_0, \quad \mathbb{E} \left[ h(X^{(n)})^{2+\delta} \right] \leq C.$$

The first assumption limits the influence of the reference trajectory in the conditional particle filter. One can check that the second assumption holds for independent and index-coupled resampling schemes. The second part of Assumption 2 ensures that if two reference trajectories are equal, an application of the CCPF returns two identical trajectories. The third assumption relates to the validity of the CPF to estimate  $\pi(h)$ , addressed under general assumptions in [Chopin and Singh \(2015\)](#); [Andrieu et al. \(2013\)](#); [Lindsten et al. \(2015\)](#). Up to the term  $\delta > 0$  which can be arbitrarily small, the assumption is a requirement if we want to estimate  $\pi(h)$  using CPFs while ensuring a finite variance.

Our main result states that the proposed estimator is unbiased and has a finite variance. Similar results can be found in e.g. Theorem 7 in [Vihola \(2015\)](#), and in [Glynn and Rhee \(2014\)](#).

**Theorem 4.1.** Under Assumptions 1-2-3, the Rhee–Glynn smoothing estimator  $H$ , given in

Eq. (1), is an unbiased estimator of  $\pi(h)$  with

$$\mathbb{E}[H^2] = \sum_{n=0}^{\infty} \mathbb{E}[(\Delta^{(n)})^2] + 2 \sum_{n=0}^{\infty} \sum_{\ell=n+1}^{\infty} \mathbb{E}[\Delta^{(n)} \Delta^{(\ell)}] < \infty,$$

where  $\Delta^{(0)} = h(X^{(0)})$  and for  $n \geq 1$ ,  $\Delta^{(n)} = h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ .

The proof is given in Appendix A. The theorem uses univariate notation for  $H$  and  $\Delta_n$ , but the Rhee–Glynn smoother can be applied to estimate multivariate smoothing functionals, for which the theorem can be interpreted component-wise.

## 5 Performance and variance reduction

The performance of the proposed estimator is tied to the meeting time. As in Chopin and Singh (2015), the coupling inequality (Lindvall, 2002) can be used to relate the meeting time with the mixing of the underlying conditional particle filter kernel. Thus, the proposed estimator is expected to work well (that is, to require few steps and to yield a small variance) in the same situations where the CPF works well (in a standard MCMC sense). It can be seen as a framework to parallelize CPF and to obtain reliable confidence intervals. Furthermore, any improvement in the CPF directly translates into a more efficient Rhee–Glynn estimator. For instance, experiments in Appendix B.2 illustrate the improvements brought by ancestor sampling, which are found to be very significant for long time series.

An important question is the scaling of the computational complexity with the time horizon. Experiments in Appendix B.3 illustrate that scaling  $N$  linearly with  $T$  yields smoothing estimators with a stable variance and a stable average meeting time; thus the overall cost of the method appears to scale quadratically in  $T$ . A theoretical study of this scaling is left for future research.

The variance of the proposed estimator can be reduced by Rao–Blackwellization. In the  $n$ -th term of the sum in Eq. (1), the random variable  $h(X^{(n)})$  is obtained by applying the test function  $h$  to a trajectory drawn among  $N$  trajectories, say  $\mathbf{x}_{0:T}$ , with probabilities  $\mathbf{w}_T$ . Thus the random variable  $\sum_{k=1}^N w_T^k h(x_{0:T}^k)$  is the conditional expectation of  $h(X^{(n)})$  given  $\mathbf{x}_{0:T}$  and  $\mathbf{w}_T$ , which has the same expectation as  $h(X^{(n)})$ . Any term  $h(X^{(n)})$  or  $h(\tilde{X}^{(n)})$  in  $H$  can be replaced by similar conditional expectations. This enables the use of all the paths generated by the CPF, and not only the selected one.

A further variance reduction can be achieved in the following way. Let  $M, m$  be two integers such that  $M > m \geq 0$ . Define

$$H_{m,M} = h(X^{(m)}) + \sum_{n=m+1}^M h(X^{(n)}) - h(\tilde{X}^{(n-1)}) \quad (3)$$

$$= h(X^{(M)}) + \sum_{n=m}^{M-1} h(X^{(n)}) - h(\tilde{X}^{(n)}). \quad (4)$$

We have  $\mathbb{E}[H_{m,M}] = \mathbb{E}[h(X^{(M)})]$  by Eq. (4) and using the fact that  $X^{(n)}$  and  $\tilde{X}^{(n)}$  have the same distribution. Furthermore,  $\mathbb{E}[h(X^{(M)})]$  goes to  $\pi(h)$  as  $M \rightarrow \infty$  under Assumption 3. We consider the estimator  $H_{m,\infty}$ , which can be computed in a finite time as follows.

We run Algorithm 1 until step  $\max(\tau, m)$ . If  $\tau \leq m+1$ , from Eq. (3),  $H_{m,\infty} = h(X^{(m)})$  almost surely, since  $X^{(n)} = \tilde{X}^{(n-1)}$  for all  $n \geq m+1$ . If  $\tau > m+1$ ,  $H_{m,\infty} = h(X^{(m)}) + \sum_{n=m+1}^{\tau-1} h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ , again using Eq. (3). The estimator  $H_{m,\infty}$  is thus made of a single term with large probability if  $m$  is large enough; the computational cost is of  $\max(\tau, m)$  instead of  $\tau$  for the original estimator. The intuition is that the fewer terms there are in  $H_{m,\infty}$ , the smaller the variance.

Another question is whether we can average over various choices of  $m$ . We can compute  $\bar{H}_m = \sum_{n=0}^m \alpha_n H_{n,\infty}$  where  $\sum_{n=0}^m \alpha_n = 1$ ; this estimator is still unbiased. It follows (after some calculations) that

$$\bar{H}_m = \sum_{n=0}^m \alpha_n h(X^{(n)}) + \sum_{n=1}^{\tau-1} \beta_n (h(X^{(n)}) - h(\tilde{X}^{(n-1)})),$$

where  $\beta_n = \sum_{j=0}^{n-1 \wedge m} \alpha_j$ ; the choice of coefficients  $\alpha_{0:m}$  is left for future work.

## 6 Numerical experiments

Appendix B contains extensive numerical experiments for the smoother on a linear Gaussian model and on a nonlinear growth model. The effects of  $N$ , of  $T$  and of the use of ancestor sampling are discussed in details. Here we focus on challenging situations: Section 6.1 uses a toy example designed to be difficult, and Section 6.2 illustrates a real-world dynamical model where the transition density is intractable.

## 6.1 A hidden auto-regressive model with an unlikely observation

We consider the first example of [Ruiz and Kappen \(2016\)](#). The latent process is defined as  $x_0 \sim \mathcal{N}(0, \tau_0^2)$  and  $x_t = \eta x_{t-1} + \mathcal{N}(0, \tau^2)$ ; we take  $\tau_0 = 0.1$ ,  $\eta = 0.9$  and  $\tau = 0.1$  and consider  $T = 10$  time steps. The process is observed only at time  $T$ , where  $y_T = 1$  and we assume  $y_T \sim \mathcal{N}(x_T, \sigma^2)$ , with  $\sigma = 0.1$ . The observation  $y_T$  is unlikely under the model. Therefore the filtering distributions and the smoothing distributions have little overlap, particularly for times  $t$  close to  $T$ . Forward filtering backward smoothing and fixed-lag smoothing are bound to struggle in this setting, which is a stylized example of highly-informative observations ([Ruiz and Kappen, 2016](#); [Del Moral and Murray, 2015](#)).

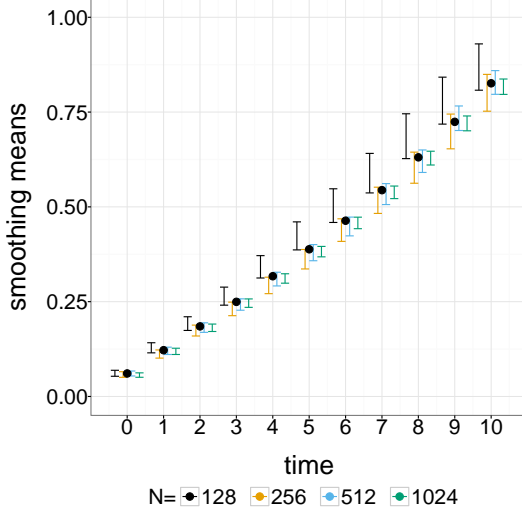
We consider the problem of estimating the smoothing means, and run  $R = 10,000$  independent Rhee–Glynn estimators, with various numbers of particles, with ancestor sampling ([Lindsten et al., 2014](#)) but without variance reduction. For comparison, we also run a bootstrap particle filter  $R$  times, with larger numbers of particles. This compensates for the fact that the Rhee–Glynn estimator requires a certain number of iterations, each involving a coupled particle filter. The average meeting times for each value of  $N$  are: 10.6 (25.1) for  $N = 128$ , 8.9 (17.0) for  $N = 256$ , 7.3 (10.8) for  $N = 512$ , 6.1 (7.3) for  $N = 1024$ .

For each method, we compute a confidence interval as  $[\hat{x}_t - 2\hat{\sigma}_t/\sqrt{R}, \hat{x}_t + 2\hat{\sigma}_t/\sqrt{R}]$  at each time  $t$ , where  $\hat{x}_t$  is the mean of the  $R$  estimators and  $\hat{\sigma}_t$  is the standard deviation. The results are shown in [Figure 1](#). The exact smoothing means are obtained analytically (black dots). The Rhee–Glynn estimators lead to reliable confidence intervals. Increasing  $N$  reduces the width of the interval and the average meeting time. On the other hand, particle filters with larger number of particles still yield unreliable confidence intervals.

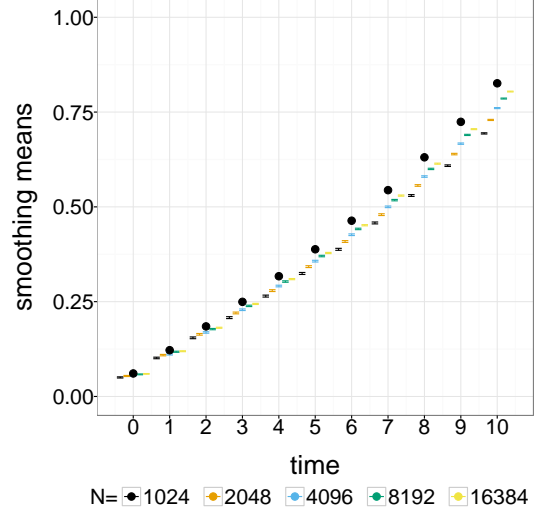
## 6.2 Prey-predator model

We investigate the performance of the Rhee–Glynn smoother for the plankton–zooplankton model of [Jones et al. \(2010\)](#), in which the transition density is intractable ([Bretó et al., 2009](#); [Jacob, 2015](#)). The hidden state  $x_t = (p_t, z_t)$  represents the population size of phytoplankton and zooplankton, and the transition from time  $t$  to  $t + 1$  is given by a Lotka–Volterra equation,

$$\frac{dp_t}{dt} = \alpha p_t - c p_t z_t, \quad \text{and} \quad \frac{dz_t}{dt} = e c p_t z_t - m_l z_t - m_q z_t^2,$$



Rhee-Glynn estimators.



Estimators obtained by particle filters.

Figure 1: Confidence intervals on the smoothing means, obtained with  $R = 10,000$  Rhee-Glynn smoothers (left), and bootstrap particle filters (right). True smoothing expectations are shown in black dots. The estimators at different times are dependent, since they are obtained from the same trajectories.

where the stochastic daily growth rate  $\alpha$  is drawn from  $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$  at every integer time  $t$ . The propagation of each particle involves solving numerically the above equation using a Runge-Kutta method in the `odeint` library (Ahnert and Mulansky, 2011). The initial distribution is given by  $\log p_0 \sim \mathcal{N}(\log 2, 1)$  and  $\log z_0 \sim \mathcal{N}(\log 2, 1)$ . The parameters  $c$  and  $e$  represent the clearance rate of the prey and the growth efficiency of the predator. Both  $m_l$  and  $m_q$  parameterize the mortality rate of the predator. The observations  $y_t$  are noisy measurements of the phytoplankton  $p_t$ ,  $\log y_t \sim \mathcal{N}(\log p_t, 0.2^2)$ ;  $z_t$  is not observed. We generate  $T = 365$  observations using  $\mu_\alpha = 0.7, \sigma_\alpha = 0.5, c = 0.25, e = 0.3, m_l = 0.1, m_q = 0.1$ . We consider the problem of estimating the mean population of zooplankton at each time  $t \in 0 : T$ , given the data-generating parameter. The intractability of the transition density precludes the use of ancestor sampling and of forward filtering backward sampling.

We draw  $R = 1,000$  independent Rhee-Glynn smoothing estimators, using  $N = 4,096$  particles. The observed meeting times have a median of 4, a mean of 4.7 and a maximum of 19. The estimator  $\hat{z}_t$  of the smoothing mean of  $z_t$  at each time  $t$  is obtained by averaging  $R = 1,000$  independent estimators. We compute the Monte Carlo variance  $\hat{v}_t$  of  $\hat{z}_t$  at each time, and define the relative variance as  $\hat{v}_t / (\hat{z}_t^2)$ .

We combine the Rhee–Glynn estimator (denoted by “unbiased” below) with Rao–Blackwellization as described in Section 5, denoted by “unbiased+RB”. Furthermore, we use the further variance reduction described in that section, denoted by “unbiased+RB+m”, which has an additional tuning parameter  $m$  chosen to be the median of the meeting time, i.e.  $m = 4$ . The latter increases the average meeting time from 4.7 to 5.1. We compare the resulting estimators with a fixed-lag smoother (Doucet and Johansen, 2011) with a lag parameter  $L = 10$ , and with a bootstrap particle filter storing the complete trajectories.

We use the same number of particles  $N = 4,096$  and compute  $R = 1,000$  estimators for each method. The relative variance is shown in Figure 2. First we see that the variance reduction techniques have a significant effect, particularly for  $t$  close to  $T$  but also for small  $t$ . The estimator  $H_{m,\infty}$  with Rao–Blackwellization (“unbiased+RB+m”) achieves nearly the same relative variance as the particle filter. The cost of these estimators can be computed as the number of iterations  $\max(m, \tau)$ , times twice the cost of a particle filter for each coupled particle filter. In the present setting where the average number of iterations is around five, we conclude that removing the bias from the standard particle filter can be done for an approximate ten-fold increase in computational cost. As expected the fixed-lag smoother leads to a significant decrease in variance. For this model, the incurred bias is negligible for  $L = 10$  (not shown), which, however, would be hard to tell if we did not have access to either unbiased methods or long runs of asymptotically exact methods.

Bootstrap particle filters and fixed-lag approximations perform well in this model. The proposed estimators are competitive in terms of variance, the tuning of the proposed algorithm is minimal, and the unbiased property prevents the possibility of over-confident error bars as in Section 6.1. Therefore the proposed method trades an extra computational cost for convenience and reliability.

## 7 Discussion

The attractive aspects of the smoother include simple parallelization and accurate error bars; these traits would be shared by perfect samplers, which aim at the more ambitious task of sampling exactly from the smoothing distribution (Lee et al., 2014).

We have shown the validity of the Rhee–Glynn estimator under mild conditions, and its behaviour as a function of the time horizon and the number of particles deserves further analysis. Numerical experiments in Appendix B investigate the effect of the time horizon and of the number of particles, as well as the significant improvements brought by ancestor sampling. Furthermore,

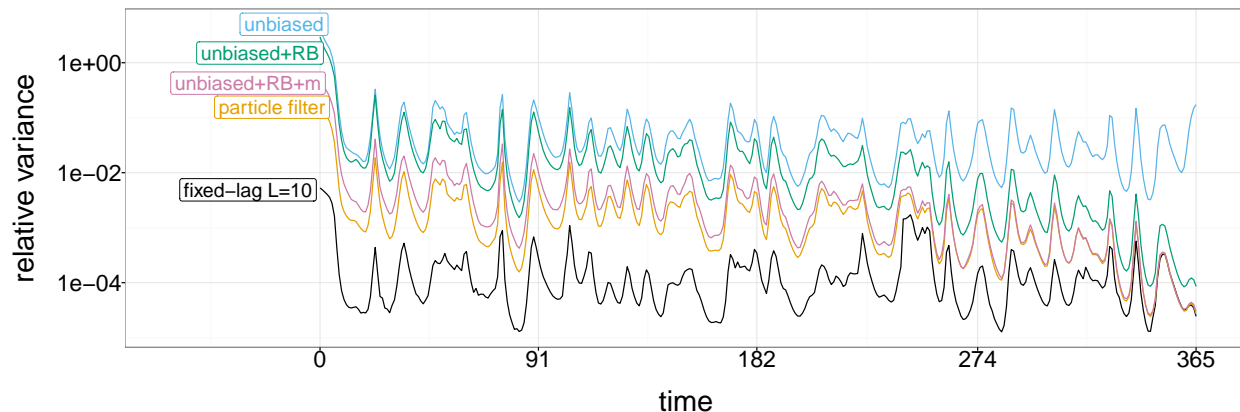


Figure 2: Comparison of the relative variance of the standard particle filter, a fixed-lag smoother with lag  $L = 10$ , and the proposed unbiased method, with Rao–Blackwellization (RB) and variance reduction (RB+m), for the estimation of the mean of the zooplankton population  $z_t$ , for the phytoplankton–zooplankton model with  $T = 365$  observations.

together with Fisher’s identity (Douc et al., 2014), the proposed smoother provides unbiased estimators of the score (for models where the transition density is tractable). This could in turn help maximizing the likelihood via stochastic gradients.

## Acknowledgements

Other applications of coupled resampling schemes have been independently proposed in Sen et al. (2016). The first author thanks Mathieu Gerber and Marco Cuturi for helpful discussions. This work was initiated during the workshop on *Advanced Monte Carlo methods for complex inference problems* at the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK held in April 2014. We would like to thank the organizers for a great event which led to this work.

## References

- Ahnert, K. and Mulansky, M. (2011). Odeint-solving ordinary differential equations in C++. *arXiv preprint arXiv:1110.3397*.
- Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):357–385.
- Andrieu, C., Lee, A., and Vihola, M. (2013). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *arXiv preprint arXiv:1312.6432*.



- Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics*, 3(1):319–348.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer-Verlag, New York.
- Chopin, N. and Singh, S. S. (2015). On particle Gibbs sampling. *Bernoulli*, 21(3):1855–1883.
- Del Moral, P. (2004). *Feynman-Kac Formulae, Genealogical and Interacting Particle Systems with Applications*. New York: Springer-Verlag.
- Del Moral, P. and Murray, L. M. (2015). Sequential Monte Carlo with highly informative observations. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):969–997.
- Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM review*, 41(1):45–76.
- Douc, R. and Cappé, O. (2005). Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 64–69.
- Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. Chapman and Hall, New York.
- Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer-Verlag, New York.
- Doucet, A., Godsill, S. J., and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208.
- Doucet, A. and Johansen, A. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering*. Oxford, UK: Oxford University Press.
- Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.*, 51A:377–389.
- Glynn, P. W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520.

- Gordon, N., Salmond, J., and Smith, A. (1993). A novel approach to non-linear/non-Gaussian Bayesian state estimation. *IEE Proceedings on Radar and Signal Processing*, 140:107–113.
- Jacob, P. E. (2015). Sequential Bayesian inference for implicit hidden Markov models and current limitations. *ESAIM: Proceedings and Surveys*, 51:24–48.
- Jacob, P. E., Lindsten, F., and Schön, T. B. (2016). Coupling of particle filters. *arXiv preprint arXiv:1606.01156*.
- Jacob, P. E., Murray, L. M., and Rubenthaler, S. (2015). Path storage in the particle filter. *Statistics and Computing*, 25(2):487–496.
- Jasra, A., Kamatani, K., Law, K. J., and Zhou, Y. (2015). Multilevel particle filter. *arXiv preprint arXiv:1510.04977*.
- Jones, E., Parslow, J., and Murray, L. (2010). A Bayesian approach to state and parameter estimation in a phytoplankton-zooplankton model. *Australian Meteorological and Oceanographic Journal*, 59:7–16.
- Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015). On particle methods for parameter estimation in state-space models. *Statistical science*, 30(3):328–351.
- Lee, A. (2008). Towards smooth particle filters for likelihood estimation with multivariate latent variables. Master’s thesis, University of British Columbia.
- Lee, A., Doucet, A., and Łatuszyński, K. (2014). Perfect simulation using atomic regeneration with application to sequential Monte Carlo. *arXiv preprints arXiv:1407.5770*.
- Lee, A. and Whiteley, N. (2015). Forest resampling for distributed sequential Monte Carlo. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(4):230–248.
- Lindsten, F., Douc, R., and Moulines, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797.
- Lindsten, F., Jordan, M. I., and Schön, T. B. (2014). Particle Gibbs with ancestor sampling. *Journal of Machine Learning Research (JMLR)*, 15:2145–2184.
- Lindsten, F. and Schön, T. B. (2013). Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends in Machine Learning*, 6(1):1–143.

- Lindvall, T. (2002). *Lectures on the coupling method*. Courier Corporation.
- Murray, L. M., Lee, A., and Jacob, P. E. (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics*, 25(3):789–805.
- Pitt, M. K. (2002). Smooth particle filters for likelihood evaluation and maximisation. *Technical report, University of Warwick, Department of Economics*.
- Rhee, C. and Glynn, P. W. (2012). A new approach to unbiased estimation for SDE’s. In *Proceedings of the Winter Simulation Conference*, pages 17:1–17:7.
- Ruiz, H.-C. and Kappen, H. (2016). Particle smoothing for hidden diffusion processes: Adaptive path integral smoother. *arXiv preprint arXiv:1605.00278*.
- Sen, D., Thiery, A., and Jasra, A. (2016). On coupling particle filter trajectories. *arXiv preprint arXiv:1606.01016*.
- Vihola, M. (2015). Unbiased estimators and multilevel Monte Carlo. *arXiv preprint arXiv:1512.01022*.
- Whiteley, N. (2010). Comment on Particle Markov chain Monte Carlo by Andrieu, Doucet and Holenstein. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):357–385.
- Williams, D. (1991). *Probability with martingales*. Cambridge university press.

## A Validity of Rhee–Glynn smoothing estimators

We first state a result on the probability of meeting in one step of CPF.

**Lemma A.1.** Under Assumptions 1 and 2, there exists  $\varepsilon > 0$  such that

$$\forall X \in \mathbb{X}^{T+1}, \quad \forall \tilde{X} \in \mathbb{X}^{T+1}, \quad \mathbb{P}(X' = \tilde{X}' | X, \tilde{X}) \geq \varepsilon,$$

where  $(X', \tilde{X}') \sim \text{CCPF}(X, \tilde{X}, \mathbf{U})$  and  $\mathbf{U} \sim \varphi$ . Furthermore, if  $X = \tilde{X}$ , then  $X' = \tilde{X}'$  almost surely.

The constant  $\varepsilon$  depends on  $N$  and  $T$ , and on the coupled resampling scheme being used. Lemma A.1 can be used, together with the coupling inequality (Lindvall, 2002), to prove the ergodicity of the conditional particle filter kernel, which is akin to the approach of Chopin and Singh (2015). The coupling inequality states that the total variation distance between  $X^{(n)}$  and  $\tilde{X}^{(n-1)}$  is less than  $2\mathbb{P}(\tau > n)$ , where  $\tau$  is the meeting time. By assuming  $\tilde{X}^{(0)} \sim \pi$ ,  $\tilde{X}^{(n)}$  follows  $\pi$  at each step  $n$ , and we obtain a bound for the total variation distance between  $X^{(n)}$  and  $\pi$ . Using Lemma A.1, we can bound the probability  $\mathbb{P}(\tau > n)$  from above by  $(1 - \varepsilon)^n$ , as in the proof of Theorem 4.1 below. This implies that the computational cost of the proposed estimator has a finite expectation for all  $N \geq 2$  and  $T$ .

### A.1 Proof of Lemma A.1

Let  $\mathcal{F}_t$  denote the filtrations generated by the CCPF at time  $t$ . We denote by  $x_{0:t}^k$ , for  $k \in 1 : N$ , the surviving trajectories at time  $t$ . Let  $I_t \subseteq 1 : N - 1$  be the set of common particles at time  $t$  defined by  $I_t = \{j \in 1 : N - 1 : x_{0:t}^j = \tilde{x}_{0:t}^j\}$ . The meeting probability, implicitly conditioned upon the reference trajectories  $x_{0:T}$  and  $\tilde{x}_{0:T}$ , can be bounded by:

$$\begin{aligned} \mathbb{P}(x'_{0:T} = \tilde{x}'_{0:T}) &= \mathbb{E} \left[ \mathbb{1} \left( x_{0:T}^{b_T} = \tilde{x}_{0:T}^{\tilde{b}_T} \right) \right] \geq \sum_{k=1}^{N-1} \mathbb{E}[\mathbb{1}(k \in I_T) P_T^{kk}] \\ &= (N-1) \mathbb{E}[\mathbb{1}(1 \in I_T) P_T^{11}] \geq \frac{N-1}{(N\bar{g})^2} \mathbb{E}[\mathbb{1}(1 \in I_T) g_T(x_T^1) g_T(\tilde{x}_T^1)], \end{aligned} \quad (5)$$

where we have used Assumptions 1 and 2. Now, let  $\psi_t : \mathbb{X}^t \mapsto \mathbb{R}_+$  and consider

$$\mathbb{E}[\mathbb{1}(1 \in I_t) \psi_t(x_{0:t}^1) \psi_t(\tilde{x}_{0:t}^1)] = \mathbb{E}[\mathbb{1}(1 \in I_t) \psi_t(x_{0:t}^1)^2], \quad (6)$$

since the two trajectories agree on  $\{1 \in I_t\}$ . We have

$$\mathbb{1}(1 \in I_t) \geq \sum_{k=1}^{N-1} \mathbb{1}(k \in I_{t-1}) \mathbb{1}(a_{t-1}^1 = \tilde{a}_{t-1}^1 = k), \quad (7)$$

and thus

$$\begin{aligned}\mathbb{E}[\mathbb{1}(1 \in I_t) \psi_t(x_{0:t}^1)^2] &\geq \mathbb{E}\left[\sum_{k=1}^{N-1} \mathbb{1}(k \in I_{t-1}) \mathbb{E}[\mathbb{1}(a_{t-1}^1 = \tilde{a}_{t-1}^1 = k) \psi_t(x_{0:t}^1)^2 \mid \mathcal{F}_{t-1}]\right] \\ &= (N-1) \mathbb{E}[\mathbb{1}(1 \in I_{t-1}) \mathbb{E}[\mathbb{1}(a_{t-1}^1 = \tilde{a}_{t-1}^1 = 1) \psi_t(x_{0:t}^1)^2 \mid \mathcal{F}_{t-1}]].\end{aligned}\quad (8)$$

The inner conditional expectation can be computed as

$$\begin{aligned}\mathbb{E}[\mathbb{1}(a_{t-1}^1 = \tilde{a}_{t-1}^1 = 1) \psi_t(x_{0:t}^1)^2 \mid \mathcal{F}_{t-1}] &= \sum_{k,\ell=1}^N P_{t-1}^{k\ell} \mathbb{1}(k = \ell = 1) \int \psi_t((x_{0:t-1}^k, x_t))^2 f(dx_t \mid x_{t-1}^k) \\ &= P_{t-1}^{11} \int \psi_t((x_{0:t-1}^1, x_t))^2 f(dx_t \mid x_{t-1}^1) \\ &\geq \frac{g_{t-1}(x_{t-1}^1) g_{t-1}(\tilde{x}_{t-1}^1)}{(N\bar{g})^2} \left( \int \psi_t((x_{0:t-1}^1, x_t)) f(dx_t \mid x_{t-1}^1) \right)^2,\end{aligned}\quad (9)$$

where we have again used Assumptions 1 and 2. Furthermore, on  $\{1 \in I_{t-1}\}$  it holds that  $x_{0:t-1}^1 = \tilde{x}_{0:t-1}^1$  and therefore, combining Eqs. (6)–(9) we get

$$\begin{aligned}\mathbb{E}[\mathbb{1}(1 \in I_t) \psi_t(x_{0:t}^1) \psi_t(\tilde{x}_{0:t}^1)] &\geq \frac{(N-1)}{(N\bar{g})^2} \mathbb{E}\left[\mathbb{1}(1 \in I_{t-1}) g_{t-1}(x_{t-1}^1) \int \psi_t((x_{0:t-1}^1, x_t)) f(dx_t \mid x_{t-1}^1) \right. \\ &\quad \left. \times g_{t-1}(\tilde{x}_{t-1}^1) \int \psi_t((\tilde{x}_{0:t-1}^1, x_t)) f(dx_t \mid \tilde{x}_{t-1}^1) \right].\end{aligned}\quad (10)$$

Thus, if we define for  $t = 1, \dots, T-1$ ,  $\psi_t(x_{0:t}) = g_t(x_t) \int \psi_{t+1}(x_{0:t+1}) f(dx_{t+1} \mid x_t)$ , and  $\psi_T(x_{0:T}) = g_T(x_T)$ , it follows that

$$\begin{aligned}\mathbb{P}(x'_{0:T} = \tilde{x}'_{0:T}) &\geq \frac{(N-1)^T}{(N\bar{g})^{2T}} \mathbb{E}[\mathbb{1}(1 \in I_1) \psi_1(x_1^1) \psi_1(\tilde{x}_1^1)] \\ &= \frac{(N-1)^T}{(N\bar{g})^{2T}} \mathbb{E}[\psi_1(x_1^1)^2] \geq \frac{(N-1)^T}{(N\bar{g})^{2T}} Z^2 > 0,\end{aligned}$$

where  $Z > 0$  is the normalizing constant of the model, defined as  $\mathbb{E}[\prod_{t=1}^T g_t(x_t)]$  where the expectation is with respect to the distribution  $m_0(dx_0) \prod_{t=1}^T f(dx_t \mid x_{t-1})$  of the latent process  $x_{0:T}$ . For any fixed  $T$ , the bound goes to zero when  $N \rightarrow \infty$ . The proof fails to capture accurately the behaviour of  $\varepsilon$  in Lemma A.1 as a function of  $N$  and  $T$ .

## A.2 Proof of Theorem 4.1

We present a proof for a generalization of the estimator given in Eq. (1). Introduce a truncation variable  $G$ , with support on the integers  $\{0, 1, 2, \dots\}$ . Define the estimator as

$$H = \sum_{n=0}^G \frac{\Delta^{(n)}}{\mathbb{P}(G \geq n)}, \quad (11)$$

where  $\Delta^{(0)} = h(X^{(0)})$  and  $\Delta^{(n)} = h(X^{(n)}) - h(\tilde{X}^{(n-1)})$ , for  $n \geq 1$ . We consider the following assumption on the truncation variable.

**Assumption 4.** The truncation variable  $G$  is Geometric, with probability mass function  $\mathbb{P}(G = n) = (1 - p)^n p$ , with support on  $\{0, 1, 2, \dots\}$  and parameter  $p \in [0, 1)$ , chosen such that  $p < 1 - (1 - \varepsilon)^{\delta/(2+\delta)}$ , where  $\varepsilon$  is as in Lemma A.1 and  $\delta$  as in Assumption 3. Furthermore,  $G$  is independent of all the other variables used in Eq. (11).

This assumption precludes the use of a range of values of  $p$  near one. On the other hand, it does not prevent the use of values of  $p$  near 0, so that we retrieve the estimator of Eq. (1) by setting  $p = 0$ , ensuring that Assumption 4 is satisfied for all values of  $\varepsilon$  and  $\delta$ . We can first upper-bound  $\mathbb{P}(\tau > n)$ , for all  $n \geq 2$ , using Lemma A.1 (e.g. Williams, 1991), exercise E.10.5. We obtain for all  $n \geq 2$ ,

$$\mathbb{P}(\tau > n) \leq (1 - \varepsilon)^{n-1}. \quad (12)$$

This ensures that  $\mathbb{E}[\tau]$  is finite; and that  $\tau$  is almost surely finite. We then introduce the random variables

$$\forall m \geq 1, \quad Z_m = \sum_{n=0}^m \frac{\Delta^{(n)} \mathbf{1}(n \leq G)}{\mathbb{P}(n \leq G)}. \quad (13)$$

Since  $\tau$  is almost surely finite, and since  $\Delta^{(n)} = 0$  for all  $n \geq \tau$ , then  $Z_m \rightarrow Z_\tau = H$  almost surely when  $m \rightarrow \infty$ . We prove that  $(Z_m)_{m \geq 1}$  is a Cauchy sequence in  $L_2$ , i.e.  $\sup_{m' \geq m} \mathbb{E}[(Z_{m'} - Z_m)^2]$  goes to 0 as  $m \rightarrow \infty$ . We write

$$(Z_{m'} - Z_m)^2 = \sum_{n=m+1}^{m'} \frac{(\Delta^{(n)})^2 \mathbf{1}(n \leq G)}{\mathbb{P}(n \leq G)^2} + 2 \sum_{n=m+1}^{m'} \sum_{\ell=n+1}^{m'} \frac{\Delta^{(n)} \Delta^{(\ell)} \mathbf{1}(\ell \leq G)}{\mathbb{P}(n \leq G) \mathbb{P}(\ell \leq G)}$$

and thus, using the independence between  $G$  and  $(\Delta^{(n)})_{n \geq 0}$ ,

$$\mathbb{E}[(Z_{m'} - Z_m)^2] = \sum_{n=m+1}^{m'} \frac{\mathbb{E}[(\Delta^{(n)})^2]}{\mathbb{P}(n \leq G)} + 2 \sum_{n=m+1}^{m'} \sum_{\ell=n+1}^{m'} \frac{\mathbb{E}[\Delta^{(n)} \Delta^{(\ell)}]}{\mathbb{P}(n \leq G)}.$$

To control  $\mathbb{E}[(\Delta^{(n)})^2] = \mathbb{E}[(\Delta^{(n)})^2 \mathbf{1}(\tau > n)]$ , we use Hölder's inequality, with  $p = 1 + \delta/2$ , and  $q = (2 + \delta)/\delta$ , where  $\delta$  is as in Assumption 3,

$$\mathbb{E}[(\Delta^{(n)})^2] \leq \mathbb{E}[(\Delta^{(n)})^{2+\delta}]^{1/(1+\delta/2)} \left( (1 - \varepsilon)^{\delta/(2+\delta)} \right)^{n-1}.$$

Furthermore, using Assumption 3, there exists  $C_1 < \infty$  such that, for  $n_0 \in \mathbb{N}$  large enough,

$$\forall n \geq n_0, \quad \mathbb{E}[(\Delta^{(n)})^{2+\delta}]^{1/(1+\delta/2)} \leq C_1. \quad (14)$$

We write  $\eta = (1 - \varepsilon)^{\delta/(2+\delta)}$ , and take  $m$  such that  $m \geq n_0$ . Using Cauchy–Schwarz, we have for all  $n, \ell \geq m$ ,

$$\mathbb{E}[\Delta^{(n)} \Delta^{(\ell)}] \leq \left( \mathbb{E}[(\Delta^{(n)})^2] \mathbb{E}[(\Delta^{(\ell)})^2] \right)^{1/2} \leq C_1 \eta^{(n-1)/2} \eta^{(\ell-1)/2}.$$

We can now write

$$\begin{aligned} \mathbb{E}[(Z_{m'} - Z_m)^2] &\leq C_1 \sum_{n=m+1}^{m'} \frac{\eta^{n-1}}{\mathbb{P}(n \leq G)} + 2 \sum_{n=m+1}^{m'} \sum_{\ell=n+1}^{m'} \frac{C_1 \eta^{(n-1)/2} \eta^{(\ell-1)/2}}{\mathbb{P}(n \leq G)} \\ &\leq C_1 \sum_{n=m+1}^{m'} \frac{\eta^{n-1}}{\mathbb{P}(n \leq G)} + 2C_1 \sum_{n=m+1}^{m'} \frac{\eta^{n-1}}{\mathbb{P}(n \leq G)} \sqrt{\eta} \frac{1 - (\sqrt{\eta})^{m'}}{1 - (\sqrt{\eta})}. \end{aligned}$$

Under Assumption 4, we have  $\mathbb{P}(n \leq G) = (1 - p)^{n+1}$ . For the above series to go to zero when  $m \rightarrow \infty$  and  $m' \geq m$ , it is enough that  $\eta/(1-p) < 1$ . By definition of  $\eta$ , this holds if  $(1 - \varepsilon)^{\delta/(2+\delta)} < 1 - p$ , which is part of Assumption 4. Thus  $(Z_m)_{m \geq 1}$  is a Cauchy sequence in  $L_2$ .

By uniqueness of the limit, since  $(Z_m)_{m \geq 1}$  goes almost surely to  $H$ ,  $(Z_m)_{m \geq 1}$  goes to  $H$  in  $L_2$ . This shows that  $H$  has finite first two moments. We can retrieve the expectation of  $H$  by

$$\mathbb{E}Z_m = \sum_{n=0}^m \mathbb{E}[\Delta^{(n)}] = \mathbb{E}[h(X^{(m)})] \xrightarrow{m \rightarrow \infty} \pi(h),$$

according to Assumption 3. We can retrieve the second moment of  $H$  by

$$\begin{aligned}\mathbb{E}[Z_m^2] &= \sum_{n=0}^m \frac{\mathbb{E}[(\Delta^{(n)})^2]}{\mathbb{P}(n \leq G)} + 2 \sum_{n=0}^m \sum_{\ell=n+1}^m \frac{\mathbb{E}[\Delta^{(n)} \Delta^{(\ell)}]}{\mathbb{P}(n \leq G)} \\ &\xrightarrow{m \rightarrow \infty} \sum_{n=0}^{\infty} \frac{\mathbb{E}[(\Delta^{(n)})^2] + 2 \sum_{\ell=n+1}^{\infty} \mathbb{E}[\Delta^{(n)} \Delta^{(\ell)}]}{\mathbb{P}(n \leq G)}.\end{aligned}$$

## B Numerical experiments

We explore the sensitivity of the proposed smoother to various inputs, section by section. The experiments are based on the hidden auto-regressive model, with  $d_x = 1$ , and the data are generated with  $\theta = 0.95$ ; except in Section B.4 where we use a nonlinear model. Each experiment is replicated  $R = 1,000$  times. We do not use any variance reduction technique in this section.

### B.1 Effect of the number of particles

We consider the effect of the number of particles  $N$ , on the meeting time and on the variance of the resulting estimator. We use a time series of length  $T = 500$ , generated from the model. As seen in the previous section, when using index-coupled resampling, we expect the meeting time  $\tau$  to occur sooner if  $N$  is larger. On the other hand, the cost of the coupled conditional particle filter is linear in  $N$ , so that the overall cost of obtaining each estimator  $H$  has expectation of order  $\mathbb{E}[\tau] \times N$ . We give estimators of this cost as a function of  $N$  in Table 1, as well as the average meeting time. We see that the cost per estimator decreases when  $N$  increases, and then increases again. There seems to be an optimal value of  $N$  yielding the minimum cost.

	cost	meeting time
N = 256	220567 (241823)	861.59 (944.62)
N = 512	17074 (17406)	33.35 (34)
N = 1024	7458 (5251)	7.28 (5.13)
N = 2048	8739 (4888)	4.27 (2.39)
N = 4096	14348 (6631)	3.5 (1.62)

Table 1: Average cost and meeting time, as a function of the number of particles  $N$ . Standard deviations are between brackets. Results obtained in the hidden auto-regressive model with  $T = 500$ .

We now consider the estimators  $H_t$  of each smoothing mean  $\mathbb{E}[x_t | y_{1:T}]$ , for  $t \in 0 : T$ , i.e. we take  $h$  to be the identity function. We compute the empirical variance of  $H_t$ , for each  $t$ , over the  $R$



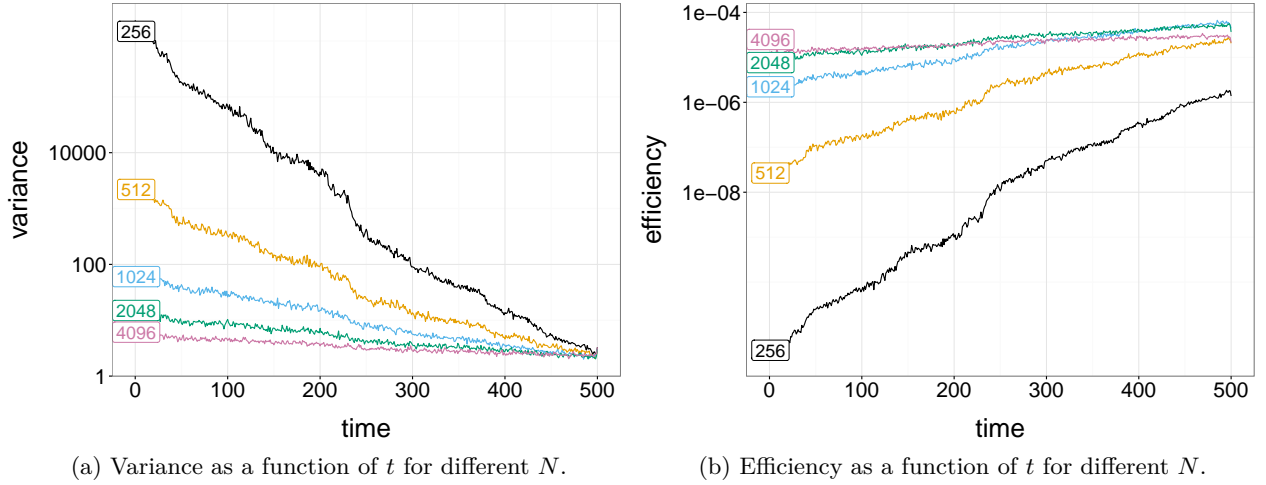


Figure 3: Variance (left) and efficiency (right) of the estimator of the smoothing mean  $\mathbb{E}[x_t|y_{1:T}]$  for  $T = 500$ , in the hidden auto-regressive model. The efficiency takes into account the computational cost of the estimator. The y-axis is on the logarithmic scale.

experiments. To take into account both variance and computational cost, we define the efficiency as  $1/(\mathbb{V}[H_t] \times \mathbb{E}[\tau] \times N)$  and approximate this value for each  $t$ , using the  $R$  estimators. The results are shown in Figure 3.

We see that the variance explodes exponentially when  $T - t$  increases (for fixed  $T$  and increasing  $t$ ; see Section B.3 for the behaviour with  $T$ ). From Figure 3a, the variance is reduced when larger values of  $N$  are used. Secondly, the variance is most reduced for the estimators of the first smoothing means, i.e.  $\mathbb{E}[x_t|y_{1:T}]$  for small  $t$ . As such, the efficiency is maximized for the largest values of  $N$  only when  $t$  is small, as can be seen from Figure 3b. For values of  $t$  closer to  $T$ , the efficiency is higher for  $N = 1,024$  and  $N = 2,048$  than it is for  $N = 4,096$ .

## B.2 Effect of ancestor sampling

We consider the use of ancestor sampling (Lindsten et al., 2014), which requires being able to evaluate the transition density,  $f(x_t|x_{t-1}, \theta)$ , for all  $x_{t-1}, x_t$  and all  $\theta$ . We set  $T = 500$  as before, and consider different values of  $N$ . The average meeting times are displayed in Table 2. We see that the meeting times are significantly reduced by using ancestor sampling, especially for smaller numbers of particles.

We consider variance and efficiency, here defined as  $1/(\mathbb{V}[H_t] \times \mathbb{E}[\tau] \times N)$ . The results are shown in Figure 4. This is to be compared with Figure 3 obtained without ancestor sampling. First we

	without ancestor sampling	with ancestor sampling
N = 256	861.59 (944.62)	8.79 (3.33)
N = 512	33.35 (34)	5.99 (2.27)
N = 1024	7.28 (5.13)	4.51 (1.88)
N = 2048	4.27 (2.39)	3.76 (1.63)
N = 4096	3.5 (1.62)	3.34 (1.51)

Table 2: Average meeting time, as a function of the number of particles  $N$ , with and without ancestor sampling. Standard deviations are between brackets. Results obtained in the hidden auto-regressive model with  $T = 500$ .

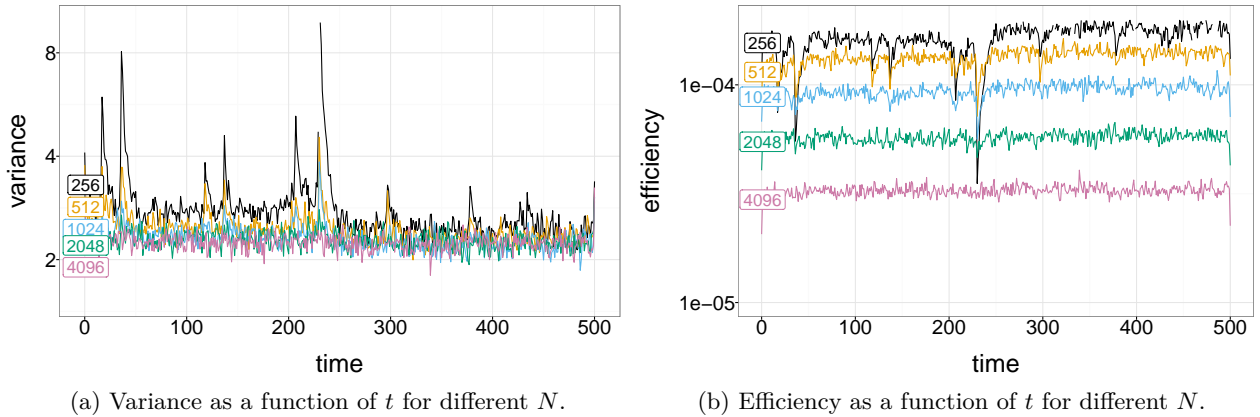


Figure 4: Variance (left) and efficiency (right) of the estimator of the smoothing mean  $\mathbb{E}[x_t|y_{1:T}]$  for  $T = 500$ , in the hidden auto-regressive model, when using ancestor sampling. The efficiency takes into account the computational cost of the estimator. The y-axis is on the logarithmic scale.

see that the variance is significantly reduced by ancestor sampling. The variance seems to increase only slowly as  $T - t$  increases, for each value of  $N$ . From Figure 4b, we see that the smallest value of  $N$  now leads to the most efficient algorithm. In other words, for a fixed computational budget, it is more efficient to produce more estimators with  $N = 256$  than to increase the number of particles and to average over fewer estimators.

### B.3 Effect of the time horizon

We investigate the effect of the time horizon  $T$ , that is, the total length of the time series, on the performance of the smoother. We expect the conditional particle filter kernel to perform less and less well when  $T$  increases. To compensate for this loss of efficiency, we increase the number of particles  $N$  linearly with  $T$ : for  $T = 64$  we use  $N = 128$ , for  $T = 128$  we use  $N = 256$ , and so forth

up to  $T = 1,024$  and  $N = 2,048$ . With that scaling, the computational cost of each run of the coupled conditional particle filter is quadratic in  $T$ . A first question is whether the meeting time is then stable with  $T$ . Table 3 reports the average meeting times obtained when scaling  $N$  linearly with  $T$ . We see that the meeting times occur in roughly the same number of steps, implying that the linear scaling of  $N$  with  $T$  is enough.

	without ancestor sampling	with ancestor sampling
$N = 128, T = 64$	11.73 (10.87)	6.54 (3.91)
$N = 256, T = 128$	9.51 (7.61)	5.77 (2.8)
$N = 512, T = 256$	11.25 (9.33)	5.66 (2.67)
$N = 1024, T = 512$	7.8 (6.05)	4.51 (1.81)
$N = 2048, T = 1024$	9.07 (6.82)	4.58 (1.9)

Table 3: Average meeting time, as a function of the number of particles  $N$  and the time horizon  $T$ , with and without ancestor sampling. Standard deviations are between brackets. Results obtained in the hidden auto-regressive model.

A second question is whether scaling  $N$  linearly with  $T$  is enough to ensure that the variance of the resulting estimator is stable. Results are shown in Figure 5, obtained without (Figure 5a) and with ancestor sampling (Figure 5b). The plots show the variance of the estimator of the smoothing means  $\mathbb{E}[x_t|y_{1:T}]$  for all  $t \leq T$  and various  $T$ . We see that, for the values of  $t$  that are less than all the time horizons, the variance of the estimators of  $\mathbb{E}[x_t|y_{1:T}]$  seems stable with  $T$ . The experiments thus indicate that, to estimate  $\mathbb{E}[x_t|y_{1:T}]$  for all  $t$ , one can scale  $N$  linearly in  $T$  and expect the meeting time and the variance of the Rhee–Glynn estimators to be stable. Overall, the computational cost is then quadratic in  $T$ .

#### B.4 Effect of multimodality in the smoothing distribution

We switch to another model to investigate the behaviour of the Rhee–Glynn estimator when the smoothing distribution is multimodal. We consider the nonlinear growth model used by Gordon et al. (1993). We set  $x_0 \sim \mathcal{N}(0, 2)$ , and, for  $t \geq 1$ ,

$$x_t = 0.5x_{t-1} + 25x_{t-1}/(1 + x_{t-1}^2) + 8\cos(1.2(t-1)) + W_t, \quad \text{and} \quad y_t = x_{t-1}^2/20 + V_t,$$

where  $W_t$  and  $V_t$  are independent normal variables, with zero means and variances 1 and 10 respectively. We generate  $T = 50$  observations using  $x_0 = 0.1$ , following Gordon et al. (1993). Because the measurement distribution  $g(y_t|x_t, \theta)$  depends on  $x_t$  through  $x_t^2$ , the sign of  $x_t$  is hard to identify,

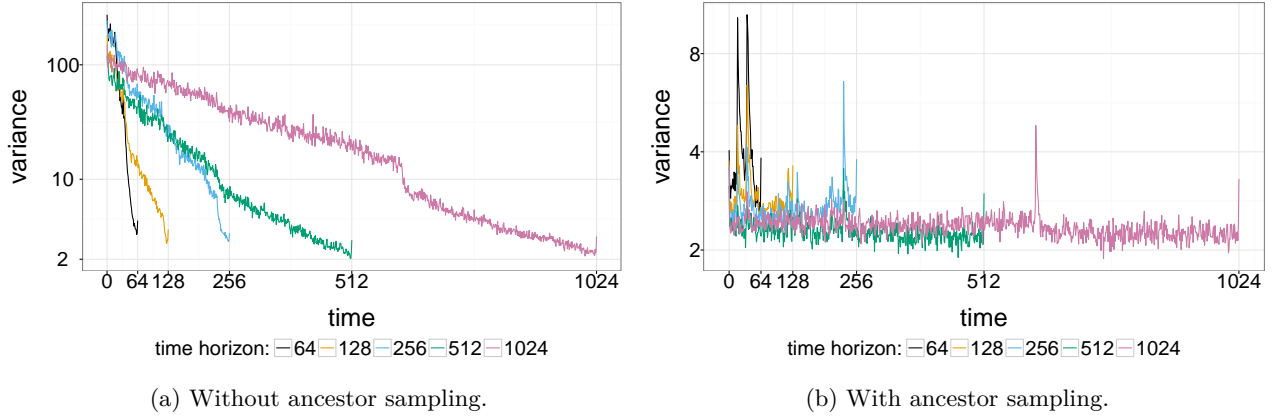
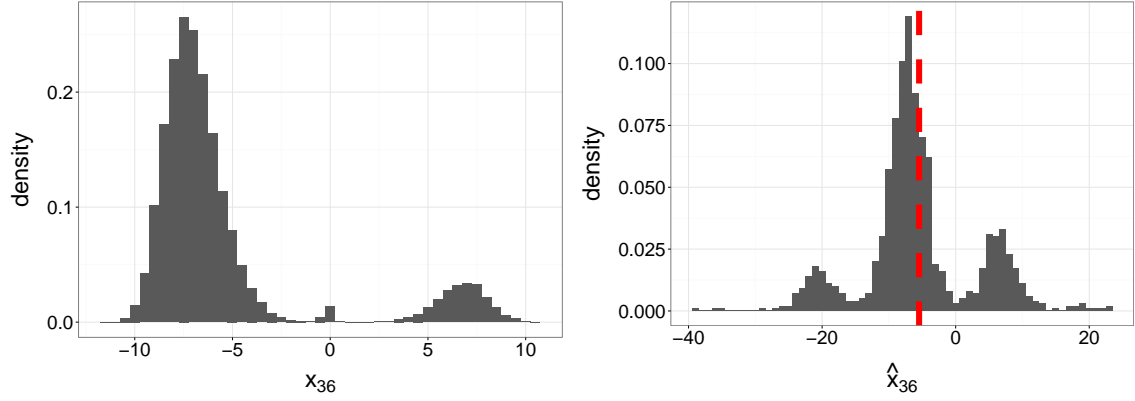


Figure 5: Variance of the estimator of the smoothing mean  $\mathbb{E}[x_t|y_{1:T}]$  for various time horizons, without (left) and with (right) ancestor sampling, in the hidden auto-regressive model. The y-axis is on the logarithmic scale.

and as a result the smoothing distribution has multiple modes. We run a conditional particle filter with ancestor sampling, with  $N = 1,024$  particles for  $M = 50,000$  iterations, and discard the first 25,000 iterations. We plot the histogram of the obtained sample for  $p(dx_t|y_{1:T}, \theta)$  at time  $t = 36$  in Figure 6a. We notice at least two modes, located around  $-7$  and  $+7$ , with possibly an extra mode near zero.

We run the Rhee–Glynn smoother with  $N = 1,024$  and ancestor sampling. Each estimator took less than 10 iterations of the coupled conditional particle filter to meet, with a median meeting time of 3 iterations. The total number of calls to the coupled conditional particle filter to obtain  $R = 1,000$  estimators adds up to 2,984. We plot the histogram of the estimators  $H_t^{(r)}$ , for  $r \in 1 : R$ , of the smoothing mean  $\mathbb{E}[x_t|y_{1:T}]$  at time  $t = 36$  in Figure 6b. We see that the distribution of the estimator is itself multimodal. Indeed, the two initial reference trajectories might belong to the mode around  $-7$ , or to the mode around  $+7$ , or each trajectory might belong to a different mode. Each of these cases leads to a mode in the distribution of the Rhee–Glynn estimator.

The resulting estimator  $\hat{x}_t$  of each smoothing mean is obtained by averaging the  $R = 1,000$  independent estimators  $H_t^{(r)}$ . We compute the Monte Carlo standard deviation  $\hat{\sigma}_t$  at each time  $t$ , and represent the confidence intervals  $[\hat{x}_t - 2\hat{\sigma}_t/\sqrt{R}, \hat{x}_t + 2\hat{\sigma}_t/\sqrt{R}]$  as error bars in Figure 7. The line represents the smoothing means obtained by conditional particle filter with ancestor sampling, taken as ground truth. The agreement shows that the proposed method is robust to multimodality in the smoothing distribution.



(a) Approximation of the smoothing distribution, using a conditional particle filter, at  $t = 36$ . (b) Rhee–Glynn estimators of the smoothing mean, and true mean in vertical dashed (red) line, at  $t = 36$ .

Figure 6: Smoothing distribution approximated by conditional particle filters (left), and  $R = 1,000$  independent Rhee–Glynn estimators of the smoothing mean (right), at time  $t = 36$  for the nonlinear growth model with  $T = 50$ .

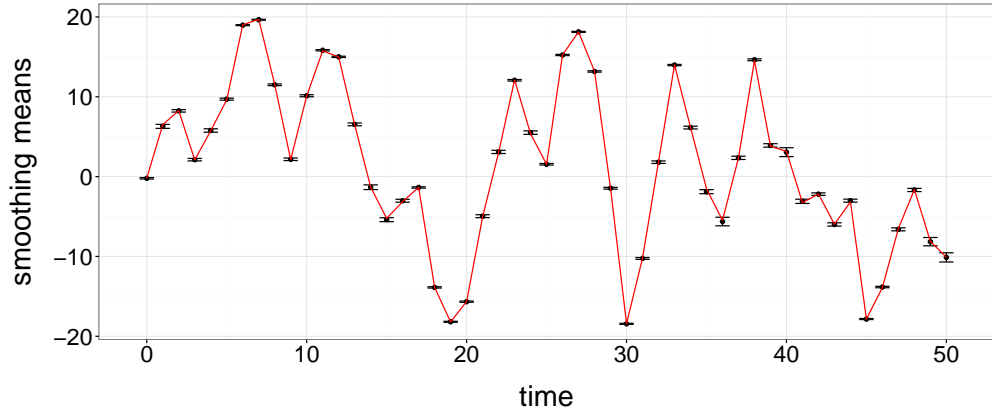


Figure 7: Confidence intervals around the exact smoothing means. The intervals are computed as two standard deviations around the mean of  $R = 1,000$  proposed smoothing estimators. The line represents the exact smoothing means, retrieved by a long run of conditional particle filter, for the nonlinear growth model with  $T = 50$  observations.

---

**Algorithm 2** Bootstrap particle filter, given a parameter  $\theta$ .

---

At step  $t = 0$ .

1. Draw  $x_0^k \sim m_0(dx_0|\theta)$ , for all  $k \in 1 : N$ .  
This can also be written  $x_0^k = M(U_0^k, \theta)$ , for all  $k \in 1 : N$ .
2. Set  $w_0^k = N^{-1}$ , for all  $k \in 1 : N$ .

At step  $t \geq 1$ .

1. Draw ancestors  $a_{t-1}^{1:N} \sim r(da^{1:N}|w_{t-1}^{1:N})$ .
  2. Draw  $x_t^k \sim f(dx_t|x_{t-1}^{a_{t-1}^k}, \theta)$ , for all  $k \in 1 : N$ .  
This can also be written  $x_t^k = F(x_{t-1}^{a_{t-1}^k}, U_t^k, \theta)$ , for all  $k \in 1 : N$ .
  3. Compute  $w_t^k \propto g(y_t|x_t^k, \theta)$ , for all  $k \in 1 : N$ , and normalize the weights.
- 

## C Pseudo-code

We provide pseudo-code for the bootstrap particle filter (Algorithm 2), the conditional particle filter (Algorithm 3), and the coupled conditional particle filter (Algorithm 4).

---

**Algorithm 3** Conditional particle filter, given a reference trajectory  $x_{0:T}$  and  $\theta$ .

---

At step  $t = 0$ .

1. Draw  $x_0^k \sim m_0(dx_0|\theta)$ , for  $k \in 1 : N - 1$ , and set  $x_0^N = x_0$ .  
This can also be written  $x_0^k = M(U_0^k, \theta)$ , for all  $k \in 1 : N - 1$ , and  $x_0^N = x_0$ .
2. Set  $w_0^k = N^{-1}$ , for  $k \in 1 : N$ .

At step  $t \geq 1$ .

1. Draw ancestors  $a_{t-1}^{1:N-1} \sim r(da^{1:N-1}|w_{t-1}^{1:N})$ , and set  $a_{t-1}^N = N$ .
2. Draw  $x_t^k \sim f(dx_t|x_{t-1}^{a_{t-1}^k}, \theta)$ , for all  $k \in 1 : N - 1$ , and set  $x_t^N = x_t$ .  
This can also be written  $x_t^k = F(x_{t-1}^{a_{t-1}^k}, U_t^k, \theta)$ , for all  $k \in 1 : N - 1$ , and  $x_t^N = x_t$ .
3. Compute  $w_t^k \propto g(y_t|x_t^k, \theta)$ , for all  $k \in 1 : N$ , and normalize the weights.

Draw a trajectory.

1. Draw  $b_T$  from a discrete distribution on  $1 : N$ , with probabilities  $w_T^{1:N}$ .
2. For  $t = T - 1, \dots, 0$ , set  $b_t = a_t^{b_{t+1}}$ .

Return  $x'_{0:T} = (x_0^{b_0}, \dots, x_T^{b_T})$ .

---

---

**Algorithm 4** Coupled conditional particle filter, given reference trajectories  $x_{0:T}$  and  $\tilde{x}_{0:T}$ .

---

At step  $t = 0$ .

1. Draw  $U_0^k$ , compute  $x_0^k = M(U_0^k, \theta)$  and  $\tilde{x}_0^k = M(U_0^k, \theta)$  for  $k \in 1 : N - 1$ .
2. Set  $x_0^N = x_0$  and  $\tilde{x}_0^N = \tilde{x}_0$ .
3. Set  $w_0^k = N^{-1}$  and  $\tilde{w}_0^k = N^{-1}$ , for  $k \in 1 : N$ .

At step  $t \geq 1$ .

1. Compute a probability matrix  $P_{t-1}$ , with marginals  $w_{t-1}^{1:N}$  and  $\tilde{w}_{t-1}^{1:N}$ . Sample  $(a_{t-1}^k, \tilde{a}_{t-1}^k)$  from  $P_{t-1}$ , for all  $k \in 1 : N - 1$ . Set  $a_{t-1}^N = N$  and  $\tilde{a}_{t-1}^N = N$ .
2. Draw  $U_t^k$ , and compute  $x_t^k = F(x_{t-1}^{a_{t-1}^k}, U_t^k, \theta)$  and  $\tilde{x}_t^k = F(\tilde{x}_{t-1}^{\tilde{a}_{t-1}^k}, U_t^k, \theta)$ , for all  $k \in 1 : N - 1$ . Set  $x_t^N = x_t$  and  $\tilde{x}_t^N = \tilde{x}_t$ .
3. Compute  $w_t^k \propto g(y_t | x_t^k, \theta)$  and  $\tilde{w}_t^k \propto g(y_t | \tilde{x}_t^k, \tilde{\theta})$ , for all  $k \in 1 : N$ , and normalize the weights.

Draw a pair of trajectories.

1. Compute a probability matrix  $P_T$ , with marginals  $w_T^{1:N}$  and  $\tilde{w}_T^{1:N}$ . Draw  $(b_T, \tilde{b}_T)$  from  $P_T$ .
2. For  $t = T - 1, \dots, 0$ , set  $b_t = a_t^{b_{t+1}}$  and  $\tilde{b}_t = \tilde{a}_t^{\tilde{b}_{t+1}}$ .

Return  $x'_{0:T} = (x_0^{b_0}, \dots, x_T^{b_T})$  and  $\tilde{x}'_{0:T} = (\tilde{x}_0^{\tilde{b}_0}, \dots, \tilde{x}_T^{\tilde{b}_T})$ .

---